To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–11

GUIDE

Identification of Genomic Markers Correlated with Sensitivity in Solid Tumors to Dasatinib Using Sparse Principal Components

Ahmed Hossain[a,*] and Hafiz T.A. Khan[b]

[a]Department of Public health, North South University, Dhaka 1229, Bangladesh.; [b]Department of Criminology & Sociology, Middlesex University, London NW4 4BT, UK.

(*v3.1 released December 2015*)

Background Differential analysis techniques are commonly used to offer scientists a dimension reduction procedure and an interpretable gateway to variable selection, especially when confronting high-dimensional genomic data. Huang *et al.* used a gene expression profile of breast cancer cell lines to identify genomic markers which are highly correlated with in vitro sensitivity of a drug Dasatinib. They considered three statistical methods to identify differentially expressed genes and finally used the results from the intersection. But the statistical methods that are used in the paper are not sufficient to select the genomic markers.
Methods: In this paper we used three alternative statistical methods to select a combined list of genomic markers and compared the genes that were proposed by Huang *et al.* We then proposed to use sparse principal component analysis (PCA) to identify a final list of genomic markers. The sparse PCA incorporates correlation into account among the genes and helps to draw a successful genomic markers discovery.
Results: We present a new and a small set of genomic markers to separate out the groups of patients effectively who are sensitive to the drug Dasatinib. The analysis procedure will also encourage scientists in identifying genomic markers that can help to separate out two groups.

Keywords: Differential gene expression; area under receiver operating characteristic curve; principal component analysis; sparse principal component analysis, clustering.

Statistics

1. Introduction

Until recently genomic study is seen to expand quite rapidly through the ongoing development of science and technologies and today this helps us to uncover many scientific questions and to understand complexities of research problems. In medical research the discovery of genomic markers opens the eyes of scientific community. It is important to identify genomic markers accurately to predict a patients response to the therapies in development. Dasatinib is a novel, oral, multi-targeted kinase inhibitor that is used for the treatment of chronic myelogenous leukemia and Philadelphia chromosome-positive acute lymphoblastic leukemia. It has been also used in clinical trials for treating patients with tumors [1].

To support the clinical development of Dasatinib, Huang *et al.* sought to identify molecular markers predictive of response to this drug that could be used for patient se-

∗ Corresponding author. Email: ahmed.hossain@utoronto.ca

lection during clinical development and beyond [2]. They used three statistical methods to identify the molecular markers: (1) signal to noise ratio (S2N), (2) Pearson correlation coefficients between expression values and another covariate (IC50), and (3) Welch t-statistic. These three statistical methods were used independently to get three lists of genes. Later, the probesets (i.e., variables) that overlapped between these three lists were considered significantly correlated with the Dasatinib sensitivity/resistance classification. Genes are found from these probsets after annotation. However, the use of these three methods may exclude highly variable genes, for which large changes in expression values can fail to enter into the list, thereby eliminating potential important biological information. Figure 1(a) presents the variances corresponding to the treatment effect (i.e. absolute mean) for the expression values of cell lines (i.e., samples) where the data is taken from the Gene Expression Omnibus (GEO)[2]. It appears that few of the probsets have low variability with high treatment effects which cause S2N or t-statistic high and increases the chance of selecting false positive genes. Figure 1(b) also confirms the presence of a number of probsets which have high variances compared to the mean (i.e., coefficient of variation in denominator). Again, correlation coefficients are highly affected by these highly variable expression values. Moreover, Huang *et al.* did not consider the correlation among the probsets instead they considered correlation coefficient between a single gene expression values and the values of IC50. These limitations in the analysis by Huang *et al.* motivate us to apply alternative approaches to prioritize genes based on three popular statistical methods and a method of sparse principal component analysis (PCA).

The statistical methods used to detect differentially expressed genes (DEGs) can be classified into two broad categories: parametric and nonparametric methods. The most commonly used parametric methods are the two-sample t-test and its variations which are based on Wald statistics. Tusher *et al.* proposed a method called significance analysis of microarray (SAM) for detecting DEGs[3]. Smyth suggested the moderated t-statistic, which generalized the sample standard deviation and are found to be robust against outliers [4, 5]. The moderated t-statistic is available in the R package `LIMMA` and the SAM is available in the R package `siggenes`. Alternatively, Hossain and Beyene considered skewed distribution instead of assuming normal distribution for expression data to identify the genomic markers [6]. All statistical tests were corrected for multiple hypotheses by using the Benjamini-Hochberg method to determine the false discovery rate (FDR) [7]. Among nonparametric methods the Wilcoxon rank sum test and area under the receiver operating characteristic curves (AUC) are widely used in gene expression analysis [8–10]. In this paper, we applied three popular methods of LIMMA, SAM and AUC.

Principal Component Analysis (PCA) is a multivariate dimension reduction and visualization technique that produces a new set of variables called principal components (PCs), constructed as linear combinations of the original variables. The limitation of PCA is that PCs are comprised of all original variables, which is unrealistic in high-dimensional genomic data and confusing, since the interpretation of PCs near impossible with large number of genomic markers. A new extension to classical PCA, sparse principal component Analysis (Sparse PCA), that systematically forces all original variables with residual contribution to have 0-valued loadings, therefore attaining a more concise and realistic group structure of the data, and a more interpretable set of PCs for further analysis [11, 12]. Hastie *et al.* proposed the popular gene shaving techniques using PCA to cluster highly variable and coherent genes in microarray datasets [13]. The lasso is a promising variable selection technique, simultaneously producing accurate and sparse models [14]. Zou and Hastie proposed the elastic net, a generalization of the lasso, which has some advantages [15]. In this paper, we applied the Sparse PCA using elastic net to summarize a gene expression profile of breast cancer cell lines (sparse PCs) for a subset

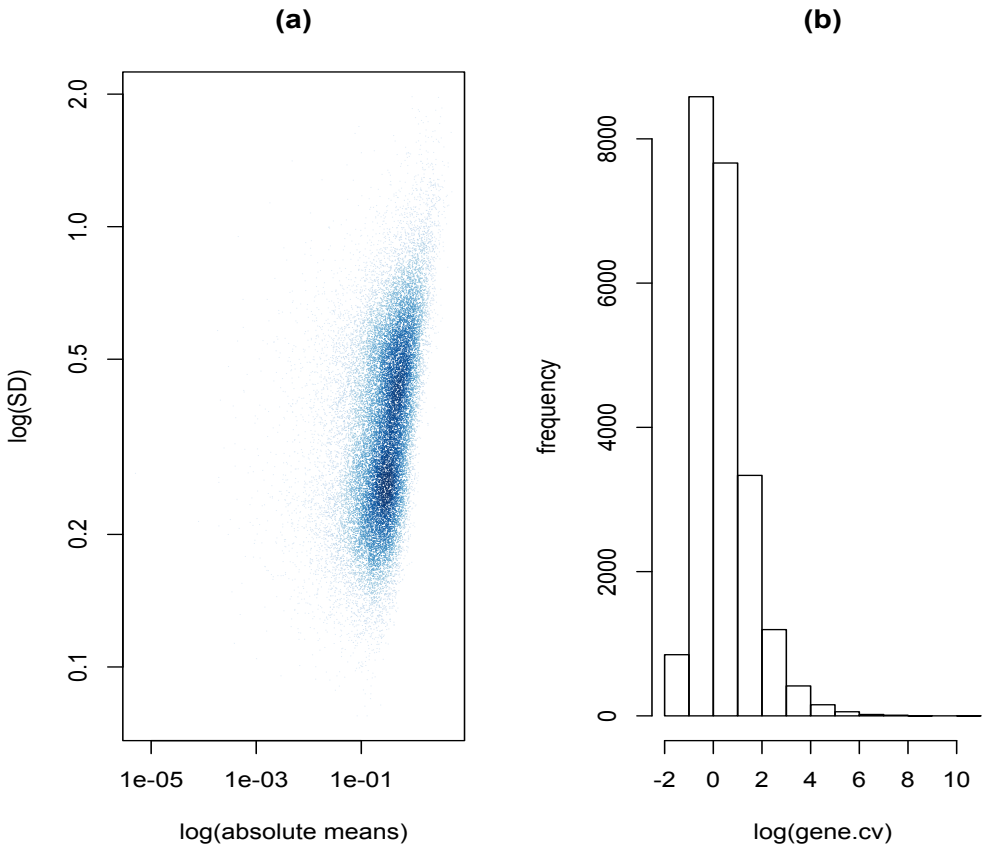

Figure 1: (a) Scatter plot presents Variability corresponding to treatment effect for each of the probsets from the Dasatinib dataset (b) Histogram for coefficient of variation (cv) presents high cvs for a number of probsets.

of the data.

2. Data and Methods

2.1 *The data*

The data is obtained from Gene Expression Omnibus (GEO), accessible through GEO series accession number GSE6569. The data contains expression of 22283 probsets (i.e., variables). Twenty-three breast cancer cell lines (i.e., samples) were used to identify candidate markers that may predict response to Dasatinib. There were 7 of them are sensitive to the drug and 16 are resistant to the drug. Thus, the matrix to be analyzed has 22283 rows of probsets and 23 columns of conditions corresponding to each sample. Details about the dataset can be found in the paper of Huang *et al.* [2]. They proposed a list of 161 genes that can be used to classify patients with sensitive and resistant to the drug Dasatinib. Starting with the 22283 probsets collected from GEO, we filtered out the bottom 75 percent in terms of Inter-Quartile Range (IQR), retaining only the 5571 most variable probsets. We removed probsets that have very little variation to begin with, since they won't provide much chance for difference detection anyway. We transformed all gene expression values using log base 2 to achieve distributions closer to normal.

2.2 *Statistical Methods*

We used three commonly used methods to identify differentially expressed genes and later we applied sparse PCA method. Perhaps uncovering an underlying genetic structure in terms of variances and correlations with a Sparse PCA method will allow for a more concise analysis. The methods are descibed briefly as follows:

2.2.1 t-test and its variation

The simplest method to detect differential gene expression is by ranking based on the fold change (FC) or ratio in expression means between the two conditions. A widely used alternative method is a t-test. The t-test is very close to signal to noise ratio that was used in the Huang *et al.* paper. Because of the large number of genes included in this experiment, there are some genes with a very small variances across cell lines (Figure 1), so that their t-values are large regardless of whether or not the differences in their averages are large. These turn out to be false positives for the t-statistic. Several alternative statistics have been proposed to overcome this problem, and many of them are influenced by the theory of shrinking the variance [3, 4, 7].

In this paper, we statistically tested marginal associations between each probset and the sensitivity/resistance to the drug Dasatinib by using moderated-t statistics (LIMMA) and SAM method. We took top 400 ranked probsets after adjusting for false-discovery rate (FDR). Huang *et al.* also considered top 400 genes considering p-values at 1% significance level. Moderated-t statistics, p-values, and FDR-adjusted p-values were calculated using the LIMMA package in R v3.26.0 [16]. SAM statistic was calculated by using `siggenes` Bioconductor package [17].

2.2.2 Area under Receiver Operating Characteristic Curve

Troyanskaya *et al.* applied the Wilcoxon rank sum test (RST) to gene expression analysis [24]. The RST is a nonparametric alternative to the two-sample t-test which is based solely on the rank of the expression values in which the observations from the two groups fall. An assessment of the expression of a gene can be made through the use of a receiver operating characteristic (ROC) curve. The ROC approach allows us considering the agreement between expression values and the presence of different thresholds simultaneously. Pepe *et al.* argue that two measures related to the ROC curve are suitable for ranking genes in regards to DE between two conditions: the Area under the ROC curve (AUC) and the partial AUC (pAUC) [8]. The AUC can be interpreted as the probability that a randomly selected subject from treatment group has greater expression values than a randomly selected subject from control group. For continuous genomic data, the nonparametric ROC curve may be preferred since it passes through all observed points and provides unbiased estimates of sensitivity, specificity, and AUC in large samples [9]. We calculated the nonparametric AUC for the expression data by using the R package WLpAUC downloaded from SIGMA website (`http://beyene-sigma-lab.com/`). We considered the top 400 probsets considering AUC values greater than 0.81.

2.2.3 Sparse principal components analysis

Principal components analysis (PCA) is a popular dimension reduction technique for determining the key variables in a multidimensional data set that explains the differences in the observations, and can be used to simplify the analysis as well as visualization of multidimensional data sets [18, 19]. The main focus of PCA is to investigate data patterns through the variance-covariance structure. The only downfall from a regression standpoint is the interpretation. For a particular PC (linear combination), the loadings

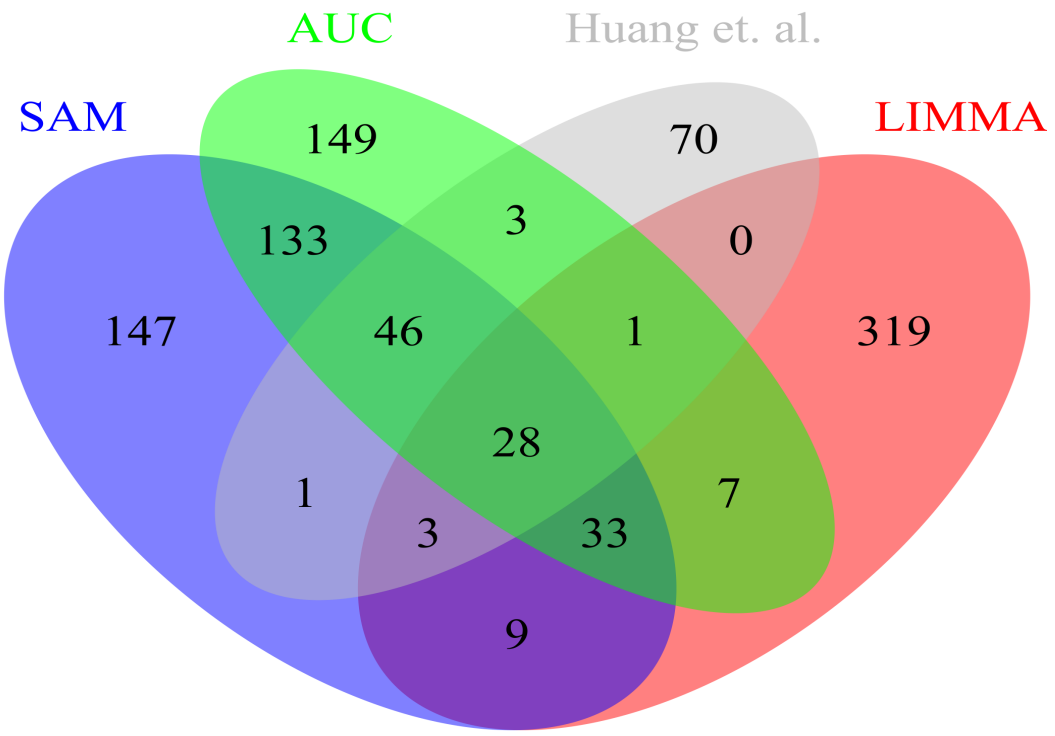

Figure 2: Venn diagram of the top ranked genes by SAM, LIMMA, AUC and Huang *et al.*. It appears that 28 genes are commonly found by these methods.

(coefficients) represent the contribution of each original variable but if there are a large number of variables, it could be very hard to determine exactly what the PC represents. The application of PCA to a genomic data doesn't allow summarizing the ways in which gene expressions vary under two biological conditions. The variance accounted for by each of the components is its associated eigen value; it is the variance of a component over all genes. Consequently, the eigenvectors with large eigen values are the ones that contain most of the information; eigen vectors with small eigen values are uninformative [19]. PCA suffers from the fact that each principal component is a linear combination of all the original variables, thus it is often difficult to interpret the results. In this circumstance we apply the sparse PCA using the lasso (or elastic net) to produce modified principal components with sparse loadings [12, 15].

3. Results

We apply the SAM, moderated-t statistic (LIMMA), and AUC and get 3 different lists of top ranked 400 probsets. After having gene annotation of these probsets we get the

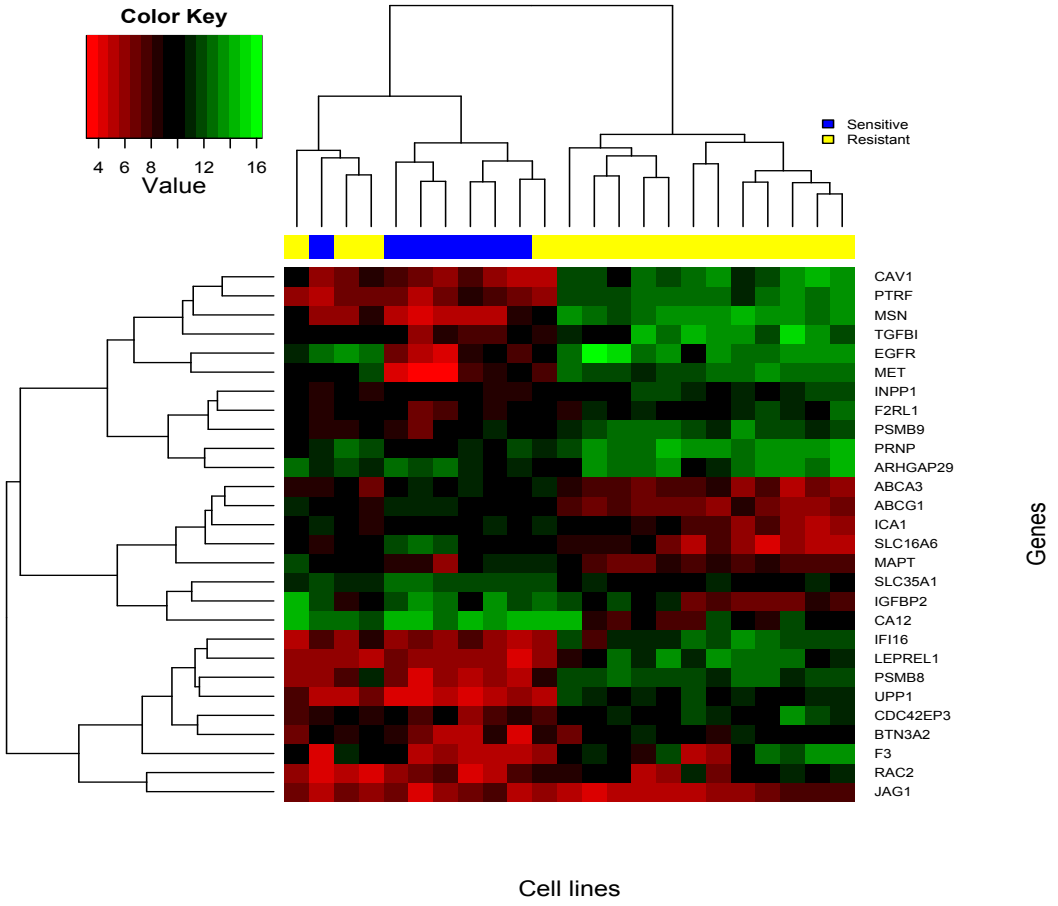

Figure 3: Cluster analysis: Heatmap with 28 concordance genes.

gene list by each of these methods. The Venn diagram in Figure 2 consists of the top ranked genes by each of the method and the genes list from Huang *et al.* It appears that all the methods produce 28 concordance genes. Concordance genes are those which are commonly found in the gene lists by the methods. In attempt to visualize some block-structure, we use a clustering algorithm to align these genes as would be needed to draw a dendrogram [21, 22]. We apply the hierarchical clustering technique to predict the sensitive/resistance of the Dasatinib and it is presented as a heatmap in Figure 3. It appears that the 28 genes can separate the samples clearly though 4 of the cell lines are misclassified. The heatmap shows patterns of color where the cell lines and genes are associated and therefore attempt to accomplish one of the missions Sparse PCA sets out to do; find natural groupings of genes.

Moving to joint associations of genes, we build 6 sparse PCs from the correlation matrix of the Dasatinib data by using the sparse PCA method. The "spca()" function in the R-package `elasticnet` was used to get the 6 sparse PCs [15]. We defined the number of sparse loadings to be obtained as 22, 22, 22, 12, 12 and 12 for the 6 sparse PCs respectively. These numbers are approximately found after testing the penalty parameters. Therefore, we found 22 non-zero loadings in the first sparse principal component and 12 non-zero loadings in the 6th sparse principal component. The non-zero elements of the loading vectors are presented in Table 1. Of the first 6 sparse PCs we investigate, the total percentage of genes variance explained by first PC is 11.2%. This is a substantial amount considering almost 96.5% of the loadings were forced to 0, validating the ability of sparse PCA as a dimension reduction technique. As one can see, the genes from SPC1 comprise a tightly-packed, high-variance group and the probsets from SPC2 and SPC3 comprise

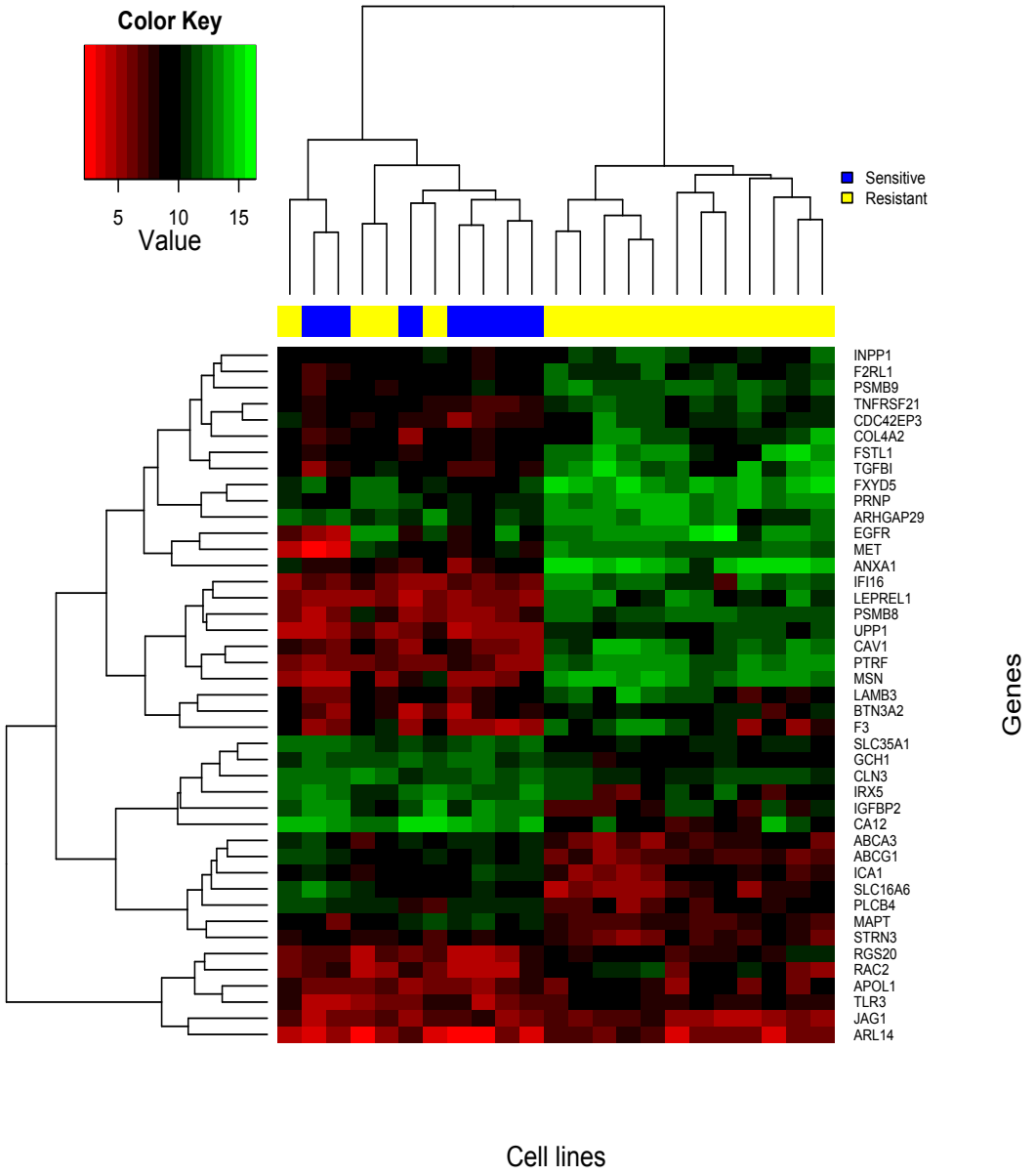

Figure 4: Cluster analysis: Heatmap with final 43 genes.

groups of lesser variance. It would now be convenient for researchers to use these SPCs to try and detect differences between sensitivity/resistent group of Dasatinib. In this regard, we fitted logistic models to the list of genes that are found in each of SPCs and estimated the prediction of a probability of a "success" (here, resistant group samples are being in case or success group). Though fitting the logistic regression model with 22 genes and 23 cell lines is a bad idea because of over-fitting problem, we did it to get an idea of comparing SPCs. These models are evaluated by receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) is calculated for each of the 6 SPCs. The results of AUC is not shown here because of over-fitting results, but we found first sparse PCs (SPC1) provide better discrimination between the resistant /sensitivity of Dasatinib after investigating the AUC values.

We found 7 concordance genes with the list of 28 genes comparing the gene list of SPC1. We suggest using these 22 genes from SPC1 because of the strength of relationship among them. Therefore we suggest a total of 43 genes, which are listed in Table 2. To further

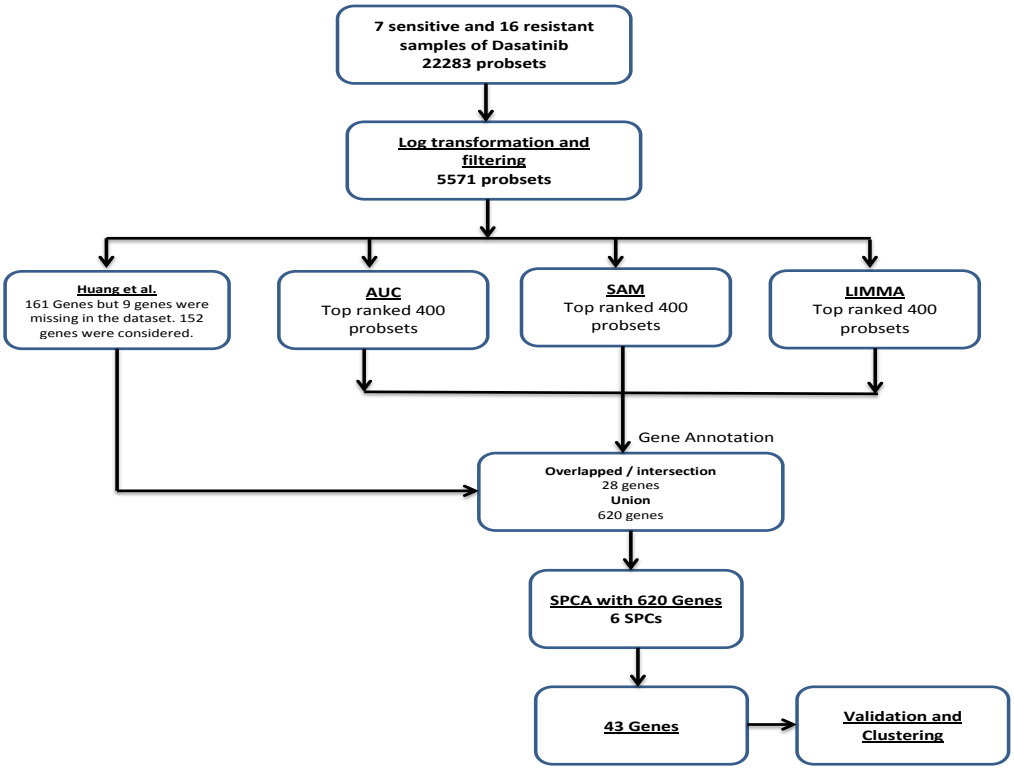

Figure 5: Flowchart of Analysis plan to get the 43 genes.

highlight potential blocks of variables, the same image strategy by a heatmap in Figure 4 is used on the sample correlation matrix and a more intensifying color scale is attempted. Though the cluster analysis is not enough to validate the results, it is necesasry to do a validation with other cell lines. It is often more beneficial to narrow down only a few genetic variants to aid researchers in further exploration; in this regard we give a very concise list of 43 genes.

In addition, we presented an analysis plan in Figure 5. The analytical procedures that applied in the study to get the 43 genes can also be used as a guideline to analyze other similar type of genomic data.

Table 1.: Genes with nonzero loadings for the first 6 sparse principal components.

	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6
1	**ABCA3**	CCDC102B	PCCA	GPR20	PARD6B	**JAG1**
2	LAMB3	TPD52	IL15RA	ELOVL2	IGF1R	BAMBI
3	**F2RL1**	ITGA5	CADM1	FRY	KCNK15	PTPN21
4	PLCB4	VAMP5	ANXA6	TAS2R13	GSDMB	SLC16A5
5	**PTRF**	KRT81	SCCPDH	KRT7	PDE4A	HLA-DQB1
6	TNFRSF21	NR2F2	CTAG2	CACNB4	ARIH2	MAOA
7	FXYD5	VGLL1	DNAJC28	DDX43	PDE4DIP	EGR2
8	STRN3	ENO2	MMRN1	KIR3DL1	IL13RA2	ZNF37BP
9	**SLC16A6**	HNRNPL	LIMA1	AP1S1	RHOBTB3	SAT1
10	FSTL1	ITPKA	LAMB1	NR5A2	RGS12	GPX1
11	**TGFBI**	RASGRF1	LINC00339	OGDHL	EIF5A	CPS1
12	IRX5	EPOR	SETDB1	MBTD1	DEFA4	TGM1
13	ARL14	KCNN4	LY6G6E			
14	ANXA1	BCL11A	TRPM2			
15	APOL1	DSC2	ENC1			
16	RGS20	ARL3	MAGEB3			
17	COL4A2	BACE2	L3MBTL1			
18	**SLC35A1**	AGAP2	ECRP			
19	CLN3	TRPM2	SGCD			
20	TLR3	TLX2	POU6F1			
21	GCH1	ARL4D	C11orf80			
22	**EGFR**	ANK3	CAST			
Concordance genes[1]	7	0	0	0	0	1
PEV[2]	11.21	8.79	4.82	4.94	4.10	2.86

[1] Concordance genes (bold) with the 28 common genes that were presented in Figure 3.
[2] Percentage of explained variance.

4. Discussion and Conclusion

Statistical genomics is a field to convert the high-dimensional data into knowledge. It is often more beneficial to narrow down only a few genetic variants to aid researchers in further exploration. The Sparse PCA method provides a new insightful way to detect important features of the data using correlation among genes into account. The study demonstrates the results of sparse PCA that helps identifying a new set of genes compared the results of Huang *et al.* We proposed a gene list that involves only a few genes, so researchers can focus on these specific genes for further analysis. The analysis helps determining which genes have changed significantly in terms of their expression between sensitivity and resistant of the Dasatinib drug. To deliver small subsets of genes, we take advantage of the computational efficiency of the three popular marginal methods and a sparse principal component analysis. The analysis also reveals classification of important groups and their associations within genes. Overall, the results have given us a better understanding in classifying groups of genomic markers that are associated with sensitivity/ resistant of Dasatinib for the breast cancer patients. However, the paper presented as a research proposal of the genomic markers that are correlated with sensitivity in solid tumors in Dasatinib and clinical applications are pending of future research.

This research is motivated by the belief that the Sparse PCA method might lead to robust results in taking correlation among genes with the Dasatinib breast cancer

cell lines data since correlation among genes are very common. We discussed the steps to use the Sparse PCA method as well as other commonly used methods to identify the differntially expressed genomic markers. Therefore, in this study we provide a new set of genes that may have more biological relevance. We anticipate that the present study will help adding scientific knowledge in medical research especially the use of the drug Dasatinib to cancer patients by taking consideration of the new set of gene list. In addition, the analytical procedures that applied in the study can also be a useful guideline to analyze other similar type of genomic data.

4.1 *Acknowledgments*

AH acknowledges North South University funding. TAK would like to acknowledge Middlesex University. We would also like to thank two anonymous reviewers and the editor for insightful comments that improved the presentation and clarity of our manuscript.

References

[1] John A., , Christopher L. *Dasatinib: A potent SRC inhibitor in clinical development for the treatment of solid tumors, Cancer Treatment Reviews* (2010), 36 (6):492-500.

[2] Huang, F., Reeves K, Han X, Fairchild C, Platero S, Wong TW, Lee F, Shaw P, Clark E. *Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: rationale for patient selection* Cancer Res. (2007) 67(5): 22262238.

[3] Tusher, V.G., Tibshirani, R., Chu, G., *Significance analysis of microarray applied to the ionizing radiation response*, PNAS (2001) 98(9):5116-5121.

[4] Smyth, G. K. *Linear models and empirical Bayes methods for assessing differential expression in microarray experiment* Statistical Applications in Genetics and Molecular Biology (2004) 3(1): Article 3.

[5] Smyth, G. K. *Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'* R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), (2005) Springer, New York, pages 397-420.

[6] Hossain, A. and Beyene, J. *Application of skew-normal distribution for detecting differential expression to microRNA data* ournal of Applied Statistics (2015) 42(3):477-491.

[7] Efron, B., Tibshirani, R., Storey, J. and Tusher, V. *Empirical Bayes Analysis of a Microarray Experiment* JASA (2001) 96:1151-1160.

[8] Pepe, M.S., Longton, G. Anderson, G.L., and Schummer, M. *Selecting Differentially Expressed Genes from Microarray Experiments* Biometrics (2003), 59:133-142.

[9] Hossain, A. and Beyene, J. *An Improved Method on Wilcoxon Rank Sum Test for Gene Selection from Microarray Experiments* Communications in Statistics - Simulation and Computation (2013) 42 (7):1563-1577.

[10] Hossain, A. and Beyene, J. *Estimation of weighted log partial area under the ROC curve and its application to MicroRNA expression data* Statistical Applications in Genetics and Molecular Biology (2013) 12(6):743-755.

[11] Zou, H., Hastie, T., Tibshirani, R. *Sparse Principal Component Analysis*, Journal of Computational and Graphical Statistics (2006), 15, 265-286.

[12] Witten, D., Tibshirani, R., Hastie, T. *A penalized matrix decomposition, with application to sparse Principal components and canonical correlation analysis* Biostatistics (2009), 10:515-534.

[13] Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., and Botstein, D. *"Gene Shaving" as a Method for Identifying Distinct Sets of Genes With Similar Expression Patterns* Genome Biology (2000) 1(2): 121.

[14] Tibshirani, R. *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological), (1996) 58, 267-288.

[15] Zou, H. and Hastie, T. *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. R package version 1.1* `http://CRAN.R-project.org/package=elasticnet`.

[16] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research (2015) 43(7), pp. e47.

[17] Schwender H. *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches R package version 1.44.0* (2012).
[18] Jolliffe, I. *Principal Component Analysis* (1986) New York: Springer Verlag .
[19] Raychaudhuri, S., Stuart, J.M., and Altman, R.B. *Principal components analysis to summarize microarray experiments: Application to sporulation time series* Pacific Symposium on Biocomputing (2000), 5:452-463.
[20] Ben-Dor, A., Shamir, R., & Yakhini, Z. *Clustering gene expression patterns*, Journal of Computational Biology (1999), 6:281-297.
[21] Buja, A., Cook, D., and Swayne, D.F. *Interactive High-Dimensional Data Visualization*, Journal of Computational and Graphical Statistics (1996), 5(1):78-99.
[22] Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X., and Somogyi, R. *Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data* (2013). Pacific Symposium on Biocomputing (1998) 3:42-53.
[23] Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. *Quantitative monitoring of gene expression patterns with a complementary DNA microarray* Science (1995) 270:467-470.
[24] Troyanskaya, O.G., Garber, M., Brown, P., Botstein, D., Altman, R.B. *Nonparamteric methods for identifying differentially expressed genes in microarray data* Bioinformatics (2002) 18(11):1454-1461.

Table 2.: 43 Genes that are highly correlated with the sensitivity/ resistant of 23 breast cancer cell lines to dasatinib

	Gene symbols	logFC[1]	P.Value[2]	FDR[3]	SAMd[4]	AUC[5]
1	JAG1	3.51	3.20e-06	0.008	1.02	0.97
2	ABCA3	-2.76	2.74e-05	0.019	-1.98	0.96
3	IFI16	4.63	4.56e-05	0.024	1.16	0.91
4	ICA1	-2.65	9.16e-05	0.032	-1.10	0.89
5	LEPREL1	4.37	9.40e-05	0.032	1.14	0.86
6	F2RL1	3.62	1.37e-04	0.042	1.19	0.95
7	CA12	-3.95	1.72e-04	0.047	-2.01	0.87
8	PTRF	4.33	1.94e-04	0.050	1.11	0.85
9	SLC16A6	-3.15	3.11e-04	0.062	-1.39	0.88
10	F3	3.88	3.62e-04	0.067	1.59	0.88
11	RAC2	2.35	4.45e-04	0.072	1.38	0.90
12	PSMB9	2.31	6.15e-04	0.084	1.13	0.91
13	PSMB8	3.70	6.34e-04	0.084	1.40	0.87
14	MSN	4.50	6.68e-04	0.086	1.06	0.86
15	CAV1	5.14	7.78e-04	0.088	1.37	0.88
16	TGFBI	3.18	8.34e-04	0.089	1.19	0.87
17	UPP1	3.45	8.86e-04	0.089	1.25	0.81
18	BTN3A2	2.46	1.08e-03	0.093	1.52	0.88
19	ABCG1	-1.71	1.11e-03	0.093	-1.05	0.94
20	CDC42EP3	2.92	2.03e-03	0.109	1.81	0.87
21	IGFBP2	-2.67	3.52e-03	0.136	-1.25	0.82
22	INPP1	1.57	3.65e-03	0.136	1.06	0.88
23	MET	3.22	4.56e-03	0.142	1.40	0.90
24	MAPT	-1.79	5.99e-03	0.165	-1.29	0.88
25	SLC35A1	-1.27	6.86e-03	0.170	-1.01	0.94
26	ARHGAP29	1.51	9.48e-03	0.190	1.35	0.84
27	PRNP	1.65	1.10e-02	0.199	1.08	0.81
28	EGFR	1.91	1.36e-02	0.219	1.55	0.82
29	LAMB3	3.25	4.67e-05	0.024	0.82	0.93
30	PLCB4	-2.61	1.44e-04	0.042	-0.72	0.92
31	TNFRSF21	2.49	2.00e-04	0.050	0.87	0.93
32	FXYD5	2.70	2.63e-04	0.061	0.95	0.94
33	STRN3	-1.88	3.06e-04	0.062	-0.76	0.95
34	FSTL1	3.06	5.80e-04	0.082	0.55	0.84
35	IRX5	-2.57	1.08e-03	0.093	-0.66	0.88
36	ARL14	2.24	1.30e-03	0.099	0.68	0.88
37	ANXA1	3.83	1.44e-03	0.100	0.66	0.81
38	APOL1	1.91	2.06e-03	0.109	0.56	0.87
39	RGS20	2.25	2.72e-03	0.128	0.42	0.86
40	COL4A2	2.33	3.15e-03	0.135	1.11	0.80
41	CLN3	-1.24	8.14e-03	0.181	-0.67	0.90
42	TLR3	1.72	1.07e-02	0.197	0.66	0.84
43	GCH1	-1.38	1.31e-02	0.215	-0.49	0.81

[1]log 2 fold change between sensitive and resistant group expressions.
[2]p-value from moderated-t statistic.
[3]p-value adjusted with Benjamini and Hochbergs method to control the False Discovery Rate.
[4] Significance analysis of microarray d statistic
[5]Area under receiver operating characteristic curve.